Research Paper | Information Technology | E-ISSN No : 2454-9916 | Volume : 3 | Issue : 5 | May 2017

A REVIEW ON TECHNIQUES FOR CLASSIFICATION OF TWITTER DATA USING BIG DATA

Hardi Thakor [1] | Chandrashekhar Dubey [2]

[1] Student, Information Technology, Parul University, Vadodara, India - 391760.

[2] Assistant Professor, Information Technology, Parul University.

ABSTRACT

Social media has become a vital part of people's life. Due to this, it generates a large amount of data that need to be processed and analyze. Some technologies were not able to handle large volume of data with storage and processing of data thus big data concept comes and handle with large data. So, there should be some mechanisms which classify unstructured data into organized form which helps user to easily access required data. Classification techniques over big data provide required data to the users from large datasets more simple way. Thus handle large amount of data used to Hadoop framework. In order to adapt these techniques for classifying Twitter data into different categories and predict the class from the unknown data. A number of issues and challenges need to be addressed, which are put forward in this paper.

KEYWORDS: Big Data, Data Mining, Classification Algorithm, Hadoop, MapReduce.

1. INTRODUCTION

Nowadays, social network becomes highly used by people. Day by day people will give more importance to a social network. Today, Facebook, Twitter, You Tube, LinkedIn, Instagram etc., are highly used with millions of people. For example, Google processes data of hundreds of Petabyte (PB), Facebook generates log data of over 10 PB per month, Baidu, a Chinese company, processes data of tens of PB, and Taobao, a subsidiary of Alibaba, generates data of tens of Terabyte (TB) for online trading per day [3]. A Huge amount of data generated with types of data like structured, semi-structured, and unstructured data. Structured data includes fixed fields of the record, numbers, date, etc. Semi-structured data include HTML, XML, PDF, doc file and unstructured data includes audios, videos, images, documents, metadata, health records, the body of e-mail messages etc. To handle the large volume of data, to analyze, stored data, processing of data applied on big data. Many applications are apply in big data like, healthcare, finance, Telecommunications, Web and Digital media etc.

Twitter is a popular micro-blogging site that allows millions of users to communicate, stay in touch, and establish connections and more. Users via Twitter can post messages called as "tweets" which are limited to 140 characters containing only text or hyperlinks. The rising popularity of social networking sites like Twitter, Facebook, LinkedIn, Tumblr etc. has produced vast resources of user-generated content. As of June 2016, Twitter reports a monthly usage of 313 million active users and 1 billion user unique visits monthly to site with embedded Tweets [1]. Twitter has witnessed a tremendous increase in the number of users recently.

In this paper, brief introduction about various classification algorithm. These algorithms are applied in big data for many areas. Hadoop is open source framework used for distributed storage and processing big data sets using MapReduce programming model. Researchers are experiments with various classification algorithms with MapReduce program and compare with other algorithm and find out accuracy of that system.

The rest of the paper is organized as follows. In the next section we discuss some classification techniques. In Section 3, An introduction of Hadoop Framework. In Section 4, Literature Review of papers. In Section 5, the paper also highlights a number of Issues and Challenges. These issues generally include the disadvantages and advantages of different classifiers. In Section 6, discuss for proposed system.

2. CLASSIFICATION TECHNIQUES

2.1 Decision Tree

The Decision tree is one of the classification techniques in which classification is done by the splitting criteria. The decision tree is a flow chart like a tree structure that classifies instances by sorting them based on the attribute (feature) values. Each and every node in a decision tree represents an attribute in an instance to be classified [7]. Decision trees are classifying instances by sorting them based on feature values. Decision tree is very time consuming when the available dataset extremely large. So overcome these problem C4.5 algorithm with MapReduce programming model is used. When available of data extremely large then C4.5 algorithm performs well in short time.

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm [4][5]. C4.5 classification is like decision trees that build a tree from root node to leaf node. Decision tree is binary tree. So tree is start from root node and also some internal node which has separated with another node. And a last node of the tree is leaf node. When construct a tree, each and every level to perform for test. Limitation of decision tree is very time consuming and not support for large dataset.

C4.5 is an extension of ID3 algorithm. ID3 algorithms select a best attribute from tree and calculate entropy and information gain. While C4.5 selects one attribute data from training data and split into samples for one class then normalized information gain. Choose an attributes from the splitting data. And last attributes with normalized information gain is chosen from decision tree. C4.5 algorithm as follows:

1. Check for base case.

2. Find best attribute best_A from samples with normalized information gain.

3. Then best_A attribute with highest information gain.

4. Split S (set) with S1, S2, S3,…,Sn with best attribute.

5. Choose a decision tree with best attribute.

2.2 Naïve Bayes

The Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model"[5]. It calculates explicit probabilities for hypothesis and it is robust to noise in input data. A naive Bayes classifier assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. Naive Bayes classifiers can be trained very efficiently in a supervised learning setting. It also called idiot's Bayes, simple Bayes, and independence Bayes. Naïve Bayes algorithm is very fast to construct and not need to iterative parameter. This means it may be rapidly applied to huge data sets. Naïve Bayes algorithm to find the conditional probability of each document to belongs to each class. The conditional probability P (C|D), where C is a classes and D is a description of objects to be classified. Given a description d of a particular object, and assign the class $argmax_c P(C=c|D=d)$.

$$argmaxc\, P(C = c|D = d) = argmaxc \frac{P(D = d|C = c)P(C = c)}{P(D = d)} \quad (1)$$

The denominator $P(D = d)$ is a normalize factor that can be ignored when determining the maximum a posteriori class, as it does not depend on the class. The key term in equation (1) is $P(D = d|C = c)$, the likelihood of the given description given the class. A Bayesian classifier estimates these likelihoods from training data, but this typically requires some additional simplifying assumptions. For instance, in an attribute-value the individual is described by a vector of values a1, . . . , an for a fixed set of attributes $A_1, \ldots, A_n$. Determining $P(D = d|C = c)$ here requires an estimate of the joint probability $P(A_1 = a_1, \ldots, An = a_n|C = c)$, abbreviated to $P(a_1, \ldots, a_n|c)$. This joint probability distribution is problematic for two reasons:

1. Its size is exponential in the number of attributes n,

2. It requires a complete training set, with several examples for each possible description. These problems vanish if we can assume that all attributes are independent given the class:

$$P(A1 = a1, \dots, An = an | C = c) = \prod_{i=1}^{n} P(Ai = ai | C = c) \quad (2)$$

This assumption is usually called the naive Bayes assumption, and a Bayesian classifier using this assumption is called the naive Bayesian classifier, often abbreviated to 'naive Bayes'. Effectively, it means that we are ignoring interactions between attributes within individuals of the same class.

2.3 K-Nearest Neighbor Algorithm

K-NN algorithm is the simplest algorithm of classification algorithm and easy to understand. The nearest neighbor (NN) rule identifies the category of unknown data point on the basis of its nearest neighbor whose class is already known [5]. K-NN in which find the k nearest of neighbor is to be considered to defined class of sample data set. As the KNN classifier requires storing the whole training set, when this is not at the redundancy of the training set to alleviate this problem [6].

In K-NN a case is classified by majority node of its nearest neighbor with the case being assigned to the class most common among its k- nearest neighbor measured by a distance function. If k is one then the case is simply assign to the class of its nearest neighbor. Training samples are stored in n-dimensional pattern space and an unknown sample of the k-NN classifier searches the pattern space for the training samples that are closest to the unknown samples. The Closeness is defined in terms of Euclidean distance, where Euclidean distance between two points, $X=(x_1, x_2, \dots, x_n)$ and $Y=(y1, y2, \dots, y_n)$ is:

$$\mathbf{d}(X,Y) = \sqrt{\sum_{i=1}^{n} (Xi - Yi)^2}$$

X and Y are two compared objects and n is their number of attributes.

KNN works as follows:

1. First to determine the parameter k this is number of nearest neighbor.
2. To calculate the distance between query data and the training sample.
3. Sort the distance and determine nearest neighbor based on the k^{th} minimal distance.
4. Collect a category of Y of the nearest distance.
5. Using simple majority of category of nearest neighbor will be prediction value for the query instance.

2.4 Support Vector Machine (SVM)

SVMs introduced in COLT-92 by Boser, Guyon & Vapnik [5]. SVM is very effective method for regression, classification and pattern recognition. It is considered as a good classifier because of its high generalization performance without need prior knowledge and input space is very high. SVM is based on the concept of decision planes that defined decision boundary and point that form the decision boundary between the classes called support vector treat as parameter [7]. The main aim of SVM is to find the best classification function to distinguish between two classes in the training data. Our main problem is that how can we represent complex data and how to exclude bogus data. Support Vector Machine is a Machine Learning tool used for classification that is based on Supervised Learning which classifies points to one of two disjoint half-spaces. Support Vector Machine is a new classification method for both linear and non-linear data [6]. Linear data can easily separate by two classes whereas non-linear data are not easily distinguished between classes.

SVM will separate data between two hyperplane. The main concept of SVM is found out best classifier between two classes. Geometrically, the margin corresponds to the shortest distance between the closest data points to the hyperplane. This geometric definition allows us to explore how to maximize the margin, so that there are infinite numbers of hyperplanes. To ensure that the maximum margin hyperplanes are actually found and SVM classifier attempts to maximize the following function with respect to $\vec{w}$ and b:

$$L_p = \tfrac{1}{2}\|\vec{w}\| - \sum_{i=1}^{t} \alpha i\, yi\left(\vec{w}.\overrightarrow{Xi} + b\right) + \sum_{i=1}^{t} \alpha i \quad (3)$$

Where, t is the number of training examples and αi, i=1,…t, are non-negative numbers and Lp is called the Lagrangian. In this equation, the vectors $\vec{w}$ and constant b define the hyperplane.

3. BIG DATA AND HADOOP FRAMEWORK

The Big Data is nothing but a data, available at heterogeneous, autonomous sources, in extreme large amount, which get updated in fractions of seconds [8].

Hadoop is a scalable, open source, fault-tolerant Virtual Grid operating system architecture for data storage and processing [10]. Hadoop is basically for storing, processing with huge dataset using commodity hardware but not for small data. The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part which is a MapReduce programming model [2].

HDFS (Hadoop Distributed File System)

HDFS cluster consists of single NameNode, that manage file system of namespace and files access by clients. In addition, numbers of DataNode, which manage storage attached to the nodes that they run on. In HDFS the files are broken into blocks and these blocks are typically large of size 64 MB or 128 MB. The blocks are stored as files on the data nodes. The blocks are replicated for reliability and typically block replication factor is 3 [9]. HDFS exposes a file system namespace and allows user data to be stored in files. A file split into one or more block and these blocks are stored into DataNode. NameNode executes file system namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to DataNodes. The DataNodes are responsible for serving read and write requests from the file system's clients. The DataNodes also perform block creation, deletion, and replication upon instruction from the NameNode. Every time DataNode send "heartbeat" message to NameNode. If there is no "heartbeat" from DataNode, the NameNode replicated that DataNode into cluster.

MapReduce

MapReduce is a programming model for processing large-scale datasets in computer clusters.

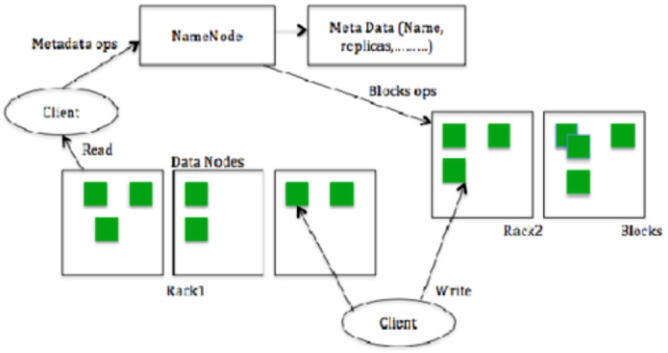

Figure 1 HDFS Architecture [9]

MapReduce is at heart of Hadoop [9]. The MapReduce programming model consists of two functions, map() and reduce(). The MapReduce function takes inputs key-value pair and produced intermediate key-value pair. In runtime system all intermediate key-value pair based on the intermediate key and passes to reduce() function.

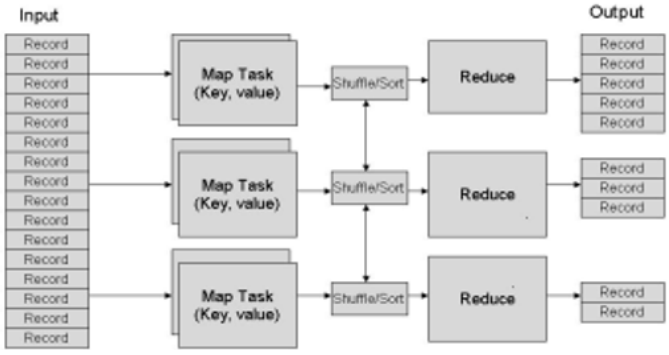

Figure 2 MapReduce Architecture and Working[10]

MapReduce working as follows shown in figure 2:

1. **Map() input:** Map processer, assigns the K1 input key and all input data associated with key-value K2.
2. **Map() code:** Map() is run exactly once for each K1 key value, generating output organized by key values K2.
3. **Shuffle and short:** Reduce processors, assigns the K2 key value each processor should work on, and provides that processor with all the Map-generated data associated with that key value. In this part combine all map tasks and produce output of intermediate key-value pair.
4. **Reduce() code:** Reduce() is run exactly once for each K2 key-value produced by the Map step.
5. **Final Output:** The MapReduce combine all the K2 key-value of reducer

part and sort by all key-value pair and generate final output.

4. LITERATURE REVIEW

M.A.Raghuram, K. Akshay and K. Chandrasekaran propose a design for a real-time system for Twitter user profile. Researcher predicts the user's static profile, tweet content and time series features of user's tweets. They present supervised machine learning approach which categorizes Twitter based on three important features (Tweets-based, User-based and Time-series based) into six categories-Politics, Entertainment, Entrepreneurship, Journalism, Science &Technology and Healthcare. They compare feature set with different classifier like Support Vector Machine, Naïve-Bayes Neighbors, Decision Tree and Logistic Regression. [11]

Jesus Maillo, Isaac Triguero, Francisco Herrera propose a MapReduce-based approach for k-Nearest neighbor classification. This model to classify large amount of text examples are against to training dataset. In the map phase will determine the k-nearest neighbor in different splits of the data. After reduce stage will compute neighbor from the map phase. They measure accuracy of correct classification and check the efficiency of parallel algorithm with slower version. They compare with sequential k-NN and MapReduce k-NN, and the result is MapReduce k-NN reduce computational time achieved by the sequential k-NN. [12]

Bayu Yudha Pratama and Riyanarto Sasrno show that using text classification to predict personality based on text written by Twitter users. Classification methods like Naive Bayes, K-Nearest Neighbors and Support Vector Machine predict the class label. After testing, the results showed Naive Bayes slightly outperformed the other methods. Analysis performs on WEKA tool. [13]

Amina Madani, Omar Boussaid and Djamel Eddine Zegour propose a new approach that discovers many different trending topics from tweets in real time. Trending topics detected for a specific geographic town and compared with the top trending topics show on twitter. Proposed the distinguished between different terms corresponding to the same trending topic. All the tweets are stored in MySQL database. Apply the generative approach of Hierarchical Dirichlet Process (HDP) for topical clustering of tweets. The system identifies new information in trending topics on Twitter content and provides meaningful analytics that give accurate description of each trending topic. [14]

Gema Bello, Hector Menendez, Shintaro Okazaki and David Camacho experiment result gives analysis about how classification and clustering techniques can interpret these opinions within a social network using information related to IKEA Company. Focused on trends extraction, similar to where Data Mining techniques are applied to extract information of users from electronic commerce. Compared with classification and clustering algorithm and then the result will show that classification is better than clustering algorithm. [15]

Prajesh P Anchalia and Kaushuk Roy implement an effective technique known as the k-Nearest Neighbor method with MapReduce program to process high volume of data. They experiments sequential k-NN vs. MapReduce for a small sized dataset. They compared MapReduce k-NN with sequential k-NN and found that MapReduce k-NN out performs the sequential k-NN with large size of datasets. [16]

Bingwei Liu, Erik Blasch, Yu Chen, Dan Shen and Genshe Chen evaluate the scalability of Naïve Bayes with MapReduce with large Datasets. They implement Naïve Bayes with MapReduce program to classify class in Positive and Negative. The result is given that the accuracy of Naïve Bayes Classifier is improved when the dataset size increases. [17]

5. ISSUES AND CHALLENGES

There are a number of techniques that can be used for classification and it is too difficult to say that a particular technique outperforms to the others. A number of issues and challenges that must be addressed to increase the overall performance and an accuracy of algorithms are discussed.

Some techniques are classification based methods to classify text on basis of labeled. The closely connected feature are extracted from the training dataset and then used to train the algorithms. Sometimes classification methods are not achieved good result. In paper [3] some issues of classification methods cannot be achieved good result. They combined classification method and then improved accuracy. Due to lack of processing capabilities and increasing number of cluster, linear speed up cannot be achieved [2]. Another issue is using the clustering technique; it is not possible to distinguish the exclusion class, which should not be neglected [5].

As per our survey, classification algorithms with MapReduce program give more accurate result and implement analysis in less time and handle large data. Hadoop Distributed File System (HDFS) and MapReduce is main component of Hadoop. HDFS is stored data with different machine and MapReduce for processing data on parallels. The performance and accuracy of classifiers depends on a number of factors.

- Accuracy is improved when features are correct. For example, the numbers of occurrences of words in many times and it is meaningless. Apart this that types of word will be removed and improve accuracy of system.
- Not all features that are returned by the tokenization algorithm must be used, because the list contains a lot of unrelated features. Thus, the feature selection method used to determines the accuracy of a classifier.
- All algorithms evaluate the keywords in a different manner and leads to different selections. Thus each algorithm required different configuration with statistical significance and number of selected features.
- The twitter domain poses a real challenge for the methods. One important point to be noted is for classify twitter data for tweets involves sentence-level classification or unlike classification of online reviews.
- Twitters users are direct their tweets to certain other users, who they follow. This is known as a Twitter mention. The general conversation is using the "@" symbol followed by the other user.
- Sometimes, external hyperlinks are included in the part of tweets and share links to the others.
- Users can also share tweets with their followers using the retweet (RT) button. RT followed by the tweet which they want to share.
- An important concept of Twitter is the hash tag. A hash tag is a part of the Tweet that denotes the topic about which the user Tweets. Normally it is preceded by"#".
- People are tweets in Twitter. So people are casual use of language also is corporate. For example, some restriction of word is replaced by, "happyyy" with happy.

The challenges is that some unwanted data are removed with existing methods have to be modified the language use on social networking websites.

6. PROPOSED SYSTEM

Our system used data collection from social networking websites, mostly Twitter for decision-making. We breakdown the process into the following subtasks:

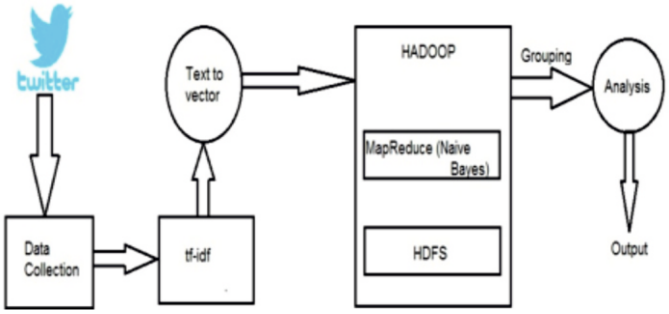

Figure 3 Block Diagram of Proposed System

Data collection:
The labeled data set of twitter users is obtained in an unstructured manner. The data is collected from the UCI repository for classification. For every Twitter user, 500 or more tweets are generated. This includes number with standard texts, special symbols, hash tags, hyperlinks etc.

Pre-processing:
The collected data is then subjected to pre-processing. Pre-processing involves filtering out non-English tweets for simplicity. Words in tweets that represent hash tags, hyperlinks, symbols and also the acronym RT (Re-tweet) is removed from tweets. Tweets that represent twitter mentions are discarded. Meaningless words like "a", "an", "the", "this" are removed from training dataset.

Feature Selection:
Tweet-based features are extracted from the tweets posted by the user. TF-IDF (Term frequency- Inverse Document Frequency) is commonly used feature in data mining. It produces a composite weight for each term in each document.

TF (t) = (Number of times term t appears in a document) / (Total number of terms in the document)

IDF (t) = log e (Total number of documents / Number of documents with term t in it).

Value (term) = TF * IDF

Analysis: The pre-processed data stored in HDFS (Hadoop Distributed File System) and classify twitter data Naïve Bayes algorithm with MapReduce program

and analysis with each tweets belongs to that particular class.

Our proposed system is classifying twitter data in different categories for example, event, health, tech, camera, art etc. These twitter data are usually in many types data but we selected only text documents and converted into vectors using the TF-IDF (term frequency- inverse document frequency) method. TF-IDF method is numeric statistics that show how important word is to a document in collection. The entire text document identifies as a vector format.

These vectors then undergo under Hadoop framework. These vectors are stored in HDFS (Hadoop Distributed File System). Naïve Bayes algorithm finds the probability of documents which belonging to the class and MapReduce preforms data processing on each server against the block of data residing on that machine. The output of the Hadoop framework is then analyzed and the result is obtained by classifying the data according to group of events.

7. CONCLUSION AND FUTURE SCOPE

Thus, it has been find out that a number of techniques can be used to perform for classify data of text. Traditional classification algorithm is good for classify data. But classification algorithm with MapReduce program gives better result. Because Hadoop framework is used to handle large amount of data, stored in HDFS (Hadoop Distributed File System) and processing with MapReduce. Techniques can be used to organize all kinds of user needs. Each technique has a different accuracy, speed and predictors.

Hence, future scope is classifying twitter data into two or more classes. And we have to use Naïve Bayes algorithm with MapReduce because algorithm is predict the class when unknown data are comes, also stored data in different machine so process are work in parallel. System defines the ability to identify the tweets belonging to a particular class.

REFERENCES

[1] Twitter: https://about.twitter.com/company

[2] Hadoop: https://en.wikipedia.org/wiki/Apache_Hadoop

[3] Chen, Min, Shiwen Mao, and Yunhao Liu, "Big data: A survey", Mobile Networks and Applications, Volume 19, Issue 2, January 2014.

[4] Kumar, Raj, and Rajesh Verma, "Classification algorithms for data mining: A survey", International Journal of Innovations in Engineering and Technology (IJIET), Vol. 1 Issue 2 August 2012.

[5] S.Archana, Dr. K.Elangovan, "Survey of Classification Techniques in Data Mining", International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 2, February- 2014.

[6] M. Sujatha, S. Prabhakar, "A Survey of Classification Techniques in Data Mining", International Journal of Innovations in Engineering and Technology (IJIET), Vol. 2 Issue 4 August 2013.

[7] Sharma, Seema, "Machine learning techniques for data mining: A survey", Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on, December 2013.

[8] Rohit Pitre, Vijay Kolekar, "A Survey Paper on Data Mining With Big Data", International Journal of Innovative Research in Advanced Engineering (IJIRAE), Volume 1 Issue 1, April 2014.

[9] Greeshma, L., and G. Pradeepini, "Big data analytics with apache hadoop mapreduce framework", Indian Journal of Science and Technology, Volume 9, Issue 26, July 2016.

[10] Manikandan, Shankar Ganesh, and Siddarth Ravi, "Big data analysis using Apache Hadoop", IT Convergence and Security (ICITCS), 2014 International Conference on. IEEE, 2014.

[11] Raghuram, M. A., K. Akshay, and K. Chandrasekaran, "Efficient User Profiling in Twitter Social Network Using Traditional Classifiers", Intelligent Systems Technologies and Applications, Vol 385 August 2015.

[12] Maillo, Jesús, Isaac Triguero, and Francisco Herrera, "A MapReduce-Based k-Nearest Neighbor Approach for Big Data Classification", Trustcom/BigDataSE/ISPA, Vol 2, December 2015.

[13] Pratama, Bayu Yudha, and Riyanarto Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM", International Conference on Data and Software Engineering March 2015.

[14] Madani, Amina, Omar Boussaid, and Djamel Eddine Zegour, "Real-time trending topics detection and description from Twitter content", Social Network Analysis and Mining, October 2015.

[15] Bello, Gema, et al, "Extracting collective trends from twitter using social-based data mining", International Conference on Computational Collective Intelligence. Vol 80832013

[16] Anchalia, Prajesh P., and Kaushik Roy, "The k-Nearest Neighbor Algorithm Using MapReduce Paradigm", 2014 5th International Conference on Intelligent Systems, Modeling and Simulation, October 2014.

[17] Liu, Bingwei, et al, "Scalable sentiment classification for big data analysis using Naive Bayes Classifier", Big Data, 2013 IEEE International Conference on, December 2013.